FITTING REDUCED RANK REGRESSION MODELS BY ALTERNATING LEAST SQUARES

Jan de Leeuw
Catrien Bijleveld
Department of Data Theory FSW
University of Leiden

Introduction

Consider the following empirical situation. For a number of **objects** or **individuals**, indexed by i = 1, ... , n, we observe two vector variables x_i and y_i. The basic idea behind this partitioning of the variables into two sets is that we have the idea that the y_i are influenced by the x_i, i.e. that the x_i are the **causes** of the y_i. The x_i can be thought of as **input variables**, the y_i as **output variables**. In econometrics the x_i are called **exogenous**, and the y_i **endogenous**. In psychometrics, and in various other areas of applied statistics, the x_i are called **independent** variables, and the y_i are **dependent**. Thus we have two **sets** of variables, and the two sets play a different and asymmetric role in our thinking.

In multivariate analysis the individuals are often considered to be replications of the same basic structure. The data can be considered to be a random sample from some well-defined population. Another way of saying this is that there is no causal connection between variables with different indices. Thus x_1 influences y_1, x_2 influences y_2, and so on, but there is no influence of x_1 on x_2 or on y_2. This is called the **independence assumption**. Another important aspect of the usual models in this class is **stationarity**. This means that the influence of x_1 of y_1 is supposed to be the same as that of x_2 on y_2, and so on. Models for which the independence assumption is violated will be discussed in a subsequent publication (De Leeuw and Bijleveld, 1987).

Independent and stationary models are at the basis of regression analysis, and of linear models in general. More recently a slightly more complicated class of these models has been discussed, which goes under various names. They are called reduced rank regression models, redundancy analysis models, growth curve models, MIMIC models, or errors-in-variables models. Their basic common idea is that the influence of x on y is mediated by unobserved **latent** variables ζ, with x determining ζ, and ζ determining y. In general the dimensionality of the ζ is lower than that of the x, and in this sense ζ **filters** the relationships between the two sets of variables. We call the space of the ζ the **latent space**, and we use p for its dimensionality. For various versions and applications of reduced rank regression we refer to Anderson (1951, 1984), Keller and Wansbeek (1983), Jøreskog and Goldberger (1975), De Leeuw, Mooijaart, and Van der Leeden (1985). The basic properties of such models will be discussed in general terms below.

In addition we consider techniques for fitting models of this kind. Some general considerations must be kept in mind here. In fitting models to data there usually are three kinds of errors that we have to take into account. The first error is **approximation error**. This occurs because models are never true, and are at best approximations. The second kind

of error is **replication error** or **sampling error**, this is the kind of error studied in statistics. It occurs because we sample from a population. It is often expedient also to discuss **measurement error**, which occurs because of limited precision or other disturbing circumstances. In survey research the measurement errors are often discussed as **non-sampling** errors. Observe that we assume that even if there are no sampling errors and no measurement errors, then there will still be approximation errors. This is because models are not exactly true, by definition. For further discussion of these points we refer to De Leeuw (1987c).

The techniques discussed in this paper are based on the general ideas of alternating least squares (Gifi, 1981, Young, 1981), or PLS (Jøreskog and Wold, 1982). They emphasize approximation errors. Or, to put it differently, we do not model sampling errors and measurement errors explicitly. Not because this is not feasible, in fact in subsequent publications we intend to include both types of errors in our modelling process. It seems preferable, however, to start with the relatively simple case in which we merely approximate complicated multivariate data structures by simpler ones. This is, in fact, the usual way in which regression and component analysis are applied in multivariate data analysis. One of the purposes of the paper is to outline the class of models, and to see how they can best be fitted using alternating least squares.

Regression with latent variables

In the usual regression situation we study the conditional distribution of y given x. This conditional distribution is studied through conditional expectations or conditional variances. Suppose p(y|x) is this conditional distribution. We use a somewhat informal notation here, which can either refer to discrete probability distribution or to densities. The purpose of statistical analysis in this context is to see if we can describe this conditional distribution in simple terms. Often this is done by assuming that the conditional expectations are linear in x, and the conditional variances do not depend on x (i.e. are homoscedastic). But this type of simplification of the models is perhaps a little bit drastic in many circumstances.

Another type of simplification can be introduced by using concepts borrowed from factor analysis. In factor analysis we observe variables $y_1,...,y_m$, and these variables are correlated. We assume that there exist unobserved variables or factors $z_1,...,z_p$ which 'explain' the association between the observed variables, in the sense that the observed variables are independent given the factors. In our informal notation we assume that

$$p(y_1,...,y_m|z) = \Pi_{j=1}^{m} p(y_j|z), \qquad (1)$$

and thus

$$p(y_1,\ldots,y_m) = \int \Pi_{j=1}^{m} p(y_j|z)\, p(z)dz. \qquad (2)$$

Now let us translate this to the regression context. The first possibility is to assume that there are p latent variables $z_1,\ldots,z_p$ such that y and x are independent given z. In a formula this is simply $p(x,y|z) = p(x|z)p(y|z)$, or, equivalently, $p(y|x,z) = p(y|z)$. It implies that

$$p(y|x) = \int p(y|z)\, p(z|x)dz. \qquad (3)$$

But the conditional independence assumption is also equivalent to $p(x|y,z) = p(x|z)$, and consequently the role played by x and y is perfectly symmetric. This is not precisely what we had in mind. We get the necessary asymmetry by assuming in addition that (1) is true. Then (3) becomes

$$p(y|x) = \int \Pi_{j=1}^{m} p(y_j|z)\, p(z|x)dz. \qquad (4)$$

Model (4) is called a **reduced rank** regression model, because z has fewer components than x. If the regressions of y on z and of z on x are linear, the name becomes even more clear. Suppose $E(y|z) = Hz$ and $E(z|x) = G'x$ then

$$E(y|x) = \int Hz\, p(z|x)dz = HG'x. \qquad (5)$$

Thus the regression coefficients B of y on x satisfy $B = HG'$, i.e. B is of reduced rank p. Observe that (5) is also true for the more general model (3).

Fitting the reduced rank regression model

Translating the models into constructive computational procedures which can be used to estimate parameters can be done in various ways. A maximum likelihood method, based on the additional assumption of joined multivariate normality, will be discussed in De Leeuw (1987). In this paper we study the method used in comparable **alternating least squares** programs for path analysis (De Leeuw, 1984a, 1987b, Coolen and De Leeuw, 1987). This

is in fact identical to the method proposed by Wold, in many different papers, for the same type of problem. See Jøreskog and Wold (1982) for an overview.

Let

$$\sigma(G,H,Z) = \mathbf{SSQ}(Z - XG) + \mathbf{SSQ}(Y - ZH'). \tag{6}$$

This choice of loss function seems natural, because the first component measures how well the latent variables Z can be predicted from the input X, and the second component measures how well the output Y can be predicted from the latent variables Z. Loss function (6) must be minimized over G, H, and Z. The general philosophy of alternating least squares leads to the idea of alternating the minimization of (6) over Z with that over G and H. Thus we start with some Z^0. We then compute G^0 and H^0 by minimizing $\sigma(G,H,Z^0)$ over G and H. This gives $G^0 = X^+Z^0$ and $H^0 = (Z^0)^+Y$, where superscript + is used for the Moore-Penrose inverse. We then compute Z^1 by minimizing $\sigma(G^0,H^0,Z)$ over Z, giving $Z^1 = (XG^0 + YH^0)[I + (H^0)'H^0]^{-1}$. Then compute G^1 and H^1, and so on. Following this procedure leads to various complications, however.

The main reason for these complications is that unconstrained minimization of (6) over G, H, and Z is not useful. This follows from Theorem 1 below. We first discuss an auxilary result. Define

$$\sigma_* = \min \mathbf{SSQ}(Y - ZH'). \tag{7}$$

In order to compute the minimum and minimizers of (7) we use familiar results on singular value decomposition, and approximation of a matrix by a matrix of given lower rank. These results are reviewed in De Leeuw (1984b). First identify the parameters by requiring Z'Z = I. This causes no loss of generality. We define $Y = K\Lambda L'$ as a singular value decomposition of Y. We truncate the decomposition at dimensionality p, retaining only the largest singular values in Λ_p, and the corresponding singular vectors in K_p and L_p. The minimum of (7) , with the identification condition Z'Z = I, is attained at $Z^\circ = K_p$ and $H^\circ = L_p\Lambda_p$. Moreover $\sigma_* = \mathbf{SSQ}(Y) - \mathbf{SSQ}(\Lambda_p^2)$.

Theorem 1. $\inf \sigma(G,H,Z) = \sigma_*$, and the infimum is attained if and only if $\text{rank}(X|YL_p) = \text{rank}(X)$.

Proof. It is clear that $\sigma(G,H,Z) \geq \sigma_*$. Now take G° arbitrary, Z° and H° as above, and define $(G,H,Z) = (\alpha G^\circ, \alpha^{-1}H^\circ, \alpha Z^\circ)$. Then $\sigma(G,H,Z) = \alpha^2 \mathbf{SSQ}(Z^\circ - XG^\circ) + \sigma_*$, and letting $\alpha \rightarrow 0$ makes $\sigma(G,H,Z) \rightarrow \sigma_*$. The minimum is attained if and only if we can choose G such that $\mathbf{SSQ}(Z^\circ - XG) = 0$, which is possible if and only if $Z^\circ = K_p = YL_p\Lambda_p^{-1}$ is in the column space of X. **QED.**

The condition in the theorem implies that the column spaces of X and Y must have a subspace of dimension at least p in common. Thus a necessary condition for the minimum to be attained is that p canonical correlations between X and Y are equal to one. In practice, of course, this means that the minimum will not be attained. Thus minimizing (6) is not a good idea, because iterative procedures will generally not converge, but produce a very large H proportional to H°, a very small Z proportional to Z°, and an arbitrary, but also very small, value of G.

In alternating least squares approaches to multivariate analysis (Gifi, 1981, Young, 1982, De Leeuw, 1983) this problem is often solved, or circumvented, by imposing a **normalization condition** on the parameters. We investigate a simple and, at first sight, natural one. It is clear that the system $Z = XG$ and $Y = ZH'$ has a solution if and only if it has a solution with $Z'Z = I$. Thus we can impose this as a normalization condition, and see what happens. It is clear that the sequence of solutions in the proof of Theorem 1 cannot be used any more, because for this sequence $Z'Z \rightarrow 0$. Moreover the normalization choice $Z'Z = I$ conforms with similar normalizations in the alternating least squares programs for component analysis and canonical analysis.

It is clear that if we normalize Z we cannot update it any more by using $Z \leftarrow (XG + YH')[I + H'H]^{-1}$. In order to find the correct update we must maximize $\text{tr } Z'(XG + YH')$ over $Z'Z = I$. This is a so-called **Procrustus problem** (Cliff, 1966), which we solve by computing $XG + YH' = K\Lambda L'$, the singular value decomposition of the matrix on the left. The optimal Z then is $Z^0 = KL'$. We use the suggestive notation $Z^0 = \mathbf{PROC}(XG + YH')$ for this. In fact from a computational point of view a somewhat simpler solution is to take $Z^0 = \mathbf{GRAM}(XG + YH')$, with **GRAM**() the Gram-Schmidt orthogonalization. While this is not optimal in the least squares sense, it only differs from the optimal Procrustus solution by a rotation. Because computing G and H in the next substep of the alternating least squares algorithm takes this rotation into account, this means that a complete step consting of **GRAM** plus G and H adjustment decreases the loss function to exactly the same value as a step consiting of **PROC** plus G and H adjustment. Since **GRAM** is simpler than **PROC** we prefer it.

An eigen analysis of the problem

The alternating least squares algorithm is mainly useful, because it can deal easily with various generalizations we will introduce at a later stage, notably the use of optimal scaling. For the problem we have outlined so far, alternating least squares is not really necessary, and probably computationally unwise. It is far easier to reformulate the problem a little bit, and to show that it can be solved by standard eigenvalue techniques. For this purpose we first project out G and H from the problem, by minimizing over them. The remaining problem is to minimize the resulting function, which is now only a function of Z, over $Z'Z = I$.

Define $\sigma(*,*,Z)$ as the minimum of (6) over G and H, for given Z satisfying $Z'Z = I$. Then

$$\sigma(*,*,Z) = p + \text{tr } Y'Y - \text{tr } Z'\{X(X'X)^{+}X' + YY'\}Z\,, \tag{8}$$

and minimizing (8) over Z satisfying $Z'Z = I$ is an eigenvalue problem. The optimal Z must satisfy the equation

$$\{X(X'X)^{+}X' + YY'\}Z = Z\Omega, \tag{9}$$

with Ω a diagonal matrix eigenvalues of the square matrix on the left. We write the optimal Z and Ω as Z_p and Ω_p, to indicate that we only use the p largest eigenvalues. Observe that X and Y enter into the calculations a bit differently, because the results are invariant under all linear transformations of the input matrix X, but not under linear transformations of the output Y. This reflects the inherent asymmetry of the problem in X and Y.

From the computational point of view it is often not very sensible to solve the eigenvalue problem (9). We can use the classical duality theory for eigenvalue problems to simplify the computations. Define $U = (\underline{X} \mid Y)$, where $\underline{X} = \textbf{GRAM}(X)$, the Gram-Schmidt orthogonalization of X, and $C = U'U$. We partition C using $C_{\underline{X}\underline{X}}$, $C_{\underline{X}Y}$, $C_{Y\underline{X}}$, C_{YY}. Thus $C_{\underline{X}\underline{X}} = I$. Now solve the eigenproblem for C, using again the p largest eigenvalues. Thus

$$C_{\underline{X}\underline{X}}A_p + C_{\underline{X}Y}B_p = A_p\Lambda_p, \tag{10a}$$

$$C_{Y\underline{X}}A_p + C_{YY}B_p = B_p\Lambda_p. \tag{10b}$$

Theorem 2. We have $\Omega_p = \Lambda_p$ and $Z_p = \mathbf{NORM}(\underset{\sim}{X}A_p + YB_p)$, where **NORM**(.) normalizes the columns of its argument to length one.

Proof. This follows directly from the singular value decomposition of U, which can be written as $U_1 = Z\Lambda^{1/2}A'$ and $U_2 = Z\Lambda^{1/2}B'$. **Q.E.D.**

We have to realize, of course, that the choice of the normalization condition $Z'Z = I$ is somewhat arbitrary. If we restrict $Z'Z$ to be equal to $\alpha^2 I$, for $\alpha^2 \neq 1$, then we can find a quite different solution. This is perhaps undesirable, because we do not want the character of the solution to be determined by the arbitrary choice of normalization. It is also unlike the situation in component analysis and canonical analysis, in which the choice of the normalization is much less essential, because different normalization conditions lead to basically the same solution.

A family of eigen solutions

Let us try to establish what the influence of the choice of α^2 in $Z'Z = \alpha^2 I$ is. Minimizing (6) under the condition that $Z'Z = \alpha^2 I$ amounts to the same thing as minimizing

$$\sigma(G,H,Z) = \alpha^2\mathbf{SSQ}(Z - XG) + \mathbf{SSQ}(Y - ZH') \qquad (11)$$

under the condition $Z'Z = I$. If α is very large, then minimizing (11) under the condition $Z'Z = I$ will amount to minimizing the second part **SSQ**(Y - ZH') of (11), under the condition that the first part **SSQ**(Z - XG) is equal to zero, i.e. under the condition that $Z = XG$ and $Z'Z = I$. We formalize this in Theorem 4, below. But first we need some additonal notation.

Suppose $\sigma(\alpha)$ is the minimum of (11) over G,H,Z, with $Z'Z = I$, and suppose $G(\alpha)$, $H(\alpha)$, and $Z(\alpha)$ are the minimizers. Using the same reasoning as before we find $H(\alpha) = Y'Z(\alpha)$ and $G(\alpha) = X^+Z(\alpha)$. Computing the optimal $Z(\alpha)$ amounts to the same thing as maximizing tr $Z'\{\alpha^2 XX^+ + YY'\}Z$ over $Z'Z = I$. Thus $Z(\alpha)$ contains eigenvectors corresponding with the p largest eigenvalues of $\alpha^2 XX^+ + YY'$.

For the limiting case we need to define

$$\sigma^\circ(G,H) = \mathbf{SSQ}(Y - XGH'), \qquad (12)$$

and $\sigma^\circ(*,*)$ is the minimum of (12) under the condition that G'X'XG = I. The minimizers are G° and H°. In order to find G° and H° we use the singular value decomposition $\underline{X}'Y = K\Lambda L'$, again with $\underline{X} = \mathbf{GRAM}(X)$, and truncate it in the usual way after p dimensions.

Lemma 3. The minimizers of (12), under the identification condition G'X'XG = I, are $G^\circ = (X'X)^{-1/2}K_p$ and $H^\circ = L_p\Lambda_p$. The minimum is $\sigma^\circ(*,*) = \mathbf{SSQ}(Y) - \mathbf{SSQ}(\Lambda_p^2)$.

Proof. In the first place (12) can be rewritten as

$$\sigma^\circ(G,H) = \mathbf{SSQ}(Y - XB_{LS}) + \mathbf{SSQ}_{X'X}(B_{LS} - GH'), \qquad (13)$$

where $B_{LS} = (X'X)^{-1}X'Y$. This shows that we are looking for the least squares rank-p approximation to $(X'X)^{1/2}B_{LS} = \underline{X}'Y$. The same classical results as before give the minimizers. From (13) we obtain for the minimum $\sigma^\circ(*,*) = \{\mathbf{SSQ}(Y) - \mathbf{SSQ}_{(X'X)^{-1}}(X'Y)\}$ $+ \{\mathbf{SSQ}_{(X'X)^{-1}}(X'Y) - \mathbf{SSQ}(\Lambda_p^2)\} = \mathbf{SSQ}(Y) - \mathbf{SSQ}(\Lambda_p^2)$. **QED.**

We now make our statement about the limiting behaviour of minimizing (11) for large α more precise.

Theorem 4. If $\alpha \rightarrow \infty$ then $\sigma(\alpha) \rightarrow \sigma^\circ(*,*)$, $G(\alpha) \rightarrow G^\circ$, $H(\alpha) \rightarrow H^\circ$, and $Z(\alpha) \rightarrow XG^\circ$.

Proof. Actually this is a special case of general results on penalty functions (Fiacco and McCormick, 1968). We shall give a proof which uses the special features of the problem (De Leeuw, 1985). We have seen that the eigenvalue problem for $\alpha^2\underline{X}\underline{X}'+ YY'$ must be solved to find $Z(\alpha)$. If α becomes large, then perturbation theory for eigenvalue problems (Kato, 1975) tells us that to find $Z(\alpha)$ we choose eigenvectors of YY' in the space spanned by X. This means that if $\alpha \rightarrow \infty$ then $Z(\alpha) \rightarrow \underline{X}M$, where $\underline{X}'YY'\underline{X}M = M\Theta$, and Θ are the p largest eigenvalues of $\underline{X}'YY'\underline{X}$. Substitution shows that Ω is approximately $\alpha^2 I + \Theta$, and $\sigma(\alpha)$ converges to tr Θ. But the elements of Θ are the same as the eigenvalues of $Y'\underline{X}\underline{X}'Y$, and thus equal to the Λ_p^2 of Lemma 3. In the same way the eigenvectors of both problems are equal. **QED.**

It is of some importance to realize that carrying out the minimization of (11) is closely related to various classical multivariate analysis techniques. This is discussed in detail in Keller and Wansbeek (1983), De Leeuw et al. (1985), or Wesselman (1987), who also give the necessary references. If X is a design matrix, then the technique becomes a form of

MANOVA, if it is an indicator matrix (the design matrix of a one-way classification) then it is canonical discriminant analysis.

Our developments so far make it possible to think of (11) as a class of criteria (or a class of normalizations). In the two extreme cases $\alpha = 0$ and $\alpha \rightarrow \infty$ we find, respectively, principal component analysis and reduced rank regression analysis (also called **redundancy analysis**, cf. Van der Wollenberg, 1977). For intermediate values of α we find intermediate solutions. It must be emphasized that these intermediate solutions are somewhat arbitrary. They depend on the normalization we choose for the latent variables Z. The only choice that seems defendable is to choose $Z'Z = I$, in the case that also diag $(X'X) = I$ and diag $(Y'Y) = I$. But the limiting case with $\alpha \rightarrow \infty$ also seems a very natural one. On the other hand using (11) with various values of α gives us a **trajectory** of solutions, starting for $\alpha = 0$ at a principal component analysis of the output, and ending for $\alpha \rightarrow \infty$ with a reduced rank analysis. It may be useful to look at this trajectory as the solution of the reduced rank regression problem. This will be illustrated in the example below.

Another remark of some importance is that if $\alpha \rightarrow \infty$, then we must have $Z = XG$ at the minimum, and the normalization $Z'Z = I$ is now used only for identification purposes. We are really minimizing (12) over G and H, and in this problem minimization is no longer needed. This is what makes the limiting case attractive. Observe that (12) can be minimized by a special alternating least squares algorithm, which alternates over G and H, no longer needing Z. This algorithm is simply $H' \leftarrow (XG)^+Y$ and $G \leftarrow X^+Y(H')^+$. It can be written more suggestively, to make it seem more relevant to (11), as $Z_1 \leftarrow XG$, $H' \leftarrow (Z_1)^+Y$, $Z_2 \leftarrow Y(H')^+$, and $G \leftarrow X^+Z_2$. This formulation does not imply, of course, that Z_1 and Z_2 converge to the same value.

An example

We analyze an example, merely for illustrative purposes, which shows what the effect in practice of our theorems is. For this purpose we take data on the fifty states of the USA, analyzed earlier by many people. We have used a version of these data taken from Meulman (1986, p. 48-54), in which there is a total of twelve variables. The first seven variables are to be considered as input variables. They are, respectively, percentage of blacks, percentage of hispanos, ratio of urban to rural, per capita income in dollars, life expectancy in years, homicide rate, and unemployment rate. The last five variables are output variables, having to do with educational achievement in the fifty states. They are: percentage high school

graduates, percentage public school enrollment, pupil teacher ratio, illiteracy rate, and failure rate on selective service mental ability test.

In order to illustrate the theorems we have first standardized all variables (sum equal to zero, sum of squares equal to one). Next we computed **GRAM**(X), and we performed the eigen-analysis of $\alpha^2\underline{XX}' + YY'$ for α equal to 0, 1, and 10. Table 1 shows the ordered eigenvalues, with the first seven eigenvalues corrected by subtracting α^2. Tables 2a, 2b, and 2c show the correlations between Z_5, the eigenvectors corresponding with the five largest eigenvalues, and the input and output variables X and Y.

Optimal Scaling

The alternating least squares techniques discussed in this paper can be combined easily with optimal scaling of the variables. This is illustrated, for example, in De Leeuw (1987b). In stead of two substeps in a main iteration, one for updating G and H for given Z and one for updating Z for given G and H, we now have three substeps. In the third substep the scaling of the variables in X and Y is updated, for given Z, G, and H.

If we take a look at loss function (6) we see that for given Z, G, and H the only part which depends on variable y_j is of the form $\mathbf{ssq}(y_j - \tilde{y}_j)$, where $\tilde{y}_j = Zh_j$. It follows that the update of variable y_j is of the form $y_j \leftarrow \mathbf{norm}(\mathbf{proj}(\tilde{y}_j))$, with **proj** denoting the projection on the cone of admissible transformations. We use **ssq** and **norm** in lower case, because they are now applied to vectors and not to matrices. The admissible transformations can be the cone of monotone transformations, the subspace of nominal transformations, a subspace of spline transformations, and so on. For details we refer to the optimal scaling literature mentioned above.

For updating variable x_i the situation is a bit more complicated. We can write the relevant part of the loss function as $\mathbf{SSQ}((Z - X_iG_i) - x_ig_i')$. Here X_iG_i contains the contributions of the input variables except x_i. Let $\tilde{x}_i = (Z - X_iG_i)g_i/\mathbf{ssq}(g_i)$. Then we have to minimize $\mathbf{ssq}(x_i - \tilde{x}_i)$, giving $x_i \leftarrow \mathbf{norm}(\mathbf{proj}(\tilde{x}_i))$. Cycling over the variables, changing them one at a time gives the third alternating least squares substep.

Of course there are many variations of this algorithm possible. We can cycle over the scaling of X and Y various times before we update Z and G and H. We can iterate the updating of Z and G and H until convergence before computing a new scaling of the variables. The general experience so far is that not too much work in each substep leads to simple computations and reasonable overall convergence, but the evidence we have for this general statement is rather shaky.

Table 1: Eigenvalues for various values of α

a:	0	1	10	100
01	2.674	2.543	2.245	2.237
02	1.331	.685	.267	.265
03	.526	.223	.166	.165
04	.323	.076	.050	.050
05	.146	.036	.028	.028
06	.000	.000	.000	.000
07	.000	.000	.000	.000
08	.000	.703	1.166	1.169
09	.000	.385	.528	.533
10	.000	.182	.269	.272
11	.000	.108	.209	.210
12	.000	.058	.071	.071

Table 2a: Loadings for $\alpha = 0$

black	.000	.000	.000	.000	.000
hispa	.000	.000	.000	.000	.000
urban	.000	.000	.000	.000	.000
incom	.000	.000	.000	.000	.000
lifex	.000	.000	.000	.000	.000
homic	.000	.000	.000	.000	.000
unemp	.000	.000	.000	.000	.000
highs	-.894	.154	.295	.400	.086
publi	-.259	.844	-.460	.086	-.047
pupil	.411	.769	.458	-.175	.002
illit	.917	-.028	.059	.320	-.230
failu	.936	.063	-.118	.150-	.289

Table 2b: Loadings for $\alpha = 1$

black	.892	-.116	-.049	-.056	-.201
hispa	.074	.403	.547	-.538	.284
urban	-.068	-.305	.537	.158	.073
incom	-.552	-.270	.608	-.013	-.359
lifex	-.742	-.231	.007	-.153	.531
homic	.763	.307	.279	.026	-.179
unemp	.277	-.152	.471	.682	.109
highs	-.839	.193	.352	-.104	-.071
publi	-.186	.802	-.281	-.118	-.042
pupil	.438	.655	.218	.233	-.014
illit	.900	-.002	.186	-.165	.118
failu	.947	-.041	.006	-.060	-.160

Table 2c: Loadings for $\alpha = 10$

black	.933	-.267	.029	-.087	-.140
hispa	.085	.642	.519	-.414	.242
urban	-.086	-.333	.586	.170	.145
incom	-.591	-.250	.655	-.013	-.303
lifex	-.786	-.156	.002	-.165	.561
homic	.806	.325	.274	.093	-.210
unemp	.279	-.221	.470	.719	.135
highs	-.768	.139	.284	-.064	-.053
publi	-.050	.376	-.224	-.079	-.038
pupil	.433	.278	.105	.167	-.020
illit	.806	.136	.150	-.098	.094
failu	.908	-.116	.045t	-.051	-.120

References

Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. **Annals of mathematical statistics**, 22, 327-351.

Anderson, T.W. (1984) Estimating linear statistical relationships. **Annals of statistics**, 12, 1-45.

Cliff, N. (1966). Orthogonal rotation to congruence. **Psychometrika,** 31, 33-42.

Coolen, H, & De Leeuw, J. (1987). **Least squares path analysis with optimal scaling.** Research Report 87-03, Department of Data Theory, University of Leiden.

De Leeuw, J. (1983) The Gifi system of nonlinear multivariate analysis. In E. Diday (ed.), **Data Analysis and Informatics IV**. Amsterdam, North Holland Publishing Company.

De Leeuw, J. (1984a). **Least squares and maximum likelihood for causal models with discrete variables.** Research Report 84-09, Department of Data Theory, University of Leiden.

De Leeuw, J. (1984b). Fixed rank approximation with singular weight matrices. **Computational Statistics Quarterly**, 1, 3-12.

De Leeuw, J. (1985). **Principal component analysis with a singular weight matrix.** Research Report 85-18. Department of Data theory, University of Leiden.

De Leeuw, J. (1987a). Nonlinear multivariate analysis with optimal scaling. In P. Legendre & L. Legendre (eds.), **Progress in Numerical Ecology**, Berlin, Springer.

De Leeuw, J. (1987b). Nonlinear path analysis with optimal scaling. In P. Legendre & L. Legendre (eds.), **Progress in Numerical Ecology**, Berlin, Springer.

De Leeuw, J. (1987c). Model choice in multinomial experiments. In T. Dijkstra (ed.), **Model choice and resampling.** Berlin, Springer.

De Leeuw, J. (1987d). **Fitting reduced rank regression models by alternating maximum likelihood estimation.** In preparation.

De Leeuw, J., & Bijleveld, C. (1987). **Fitting longitudinal reduced rank regression models by alternating least squares.** Research Report 87-00. Department of Data Theory, University of Leiden.

De Leeuw, J., Mooijaart, A, & Van der Leeden, R. (1985). **Fixed factor score models with linear restrictions.** Research Report 85-06. Department of Data Theory, University of Leiden.

A. Fiacco & G. MacCormick (1968). **Non-linear programming: Sequential Unconstrained Minimization Techniques.** New York, Wiley.

Gifi, A. (1981). **Nonlinear Multivariate Analysis.** Department of Data Theory, University of Leiden.

Jøreskog, K.G., & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. **Journal of the American Statistical Association**,70,631-639.

Jøreskog, K.G., & Wold, H.O.A. (1982). **Systems under indirect observation: causality, structure, prediction.** Amsterdam, North Holland Publishing Company.

Keller, W.J., & Wansbeek, T. (1983). Multivariate methods for quantitative and qualitative data. **Journal of Econometrics**, 22, 91-111.

Kato, T. (1975). **Perturbation theory of linear operators.** Berlin, Springer Verlag.

Meulman, J. (1986). **A distance approach to nonlinear multivariate analysis.** Leiden, DSWO-Press.

Van der Wollenberg, A.L. (1977) Redundancy analysis. An alternative for canonical analysis. **Psychometrika**, 42, 207-219.

Wesselman, A.M. (1987) **The population sample decomposition method. A distribution-free estimation technique for minimum distance parameters.** Unpublished doctoral dissertation, Erasmus University, Rotterdam.

Young, F.W. (1981) Quantitative analysis of qualitative data. **Psychometrika**, 46, 347-388.